Business Experimentation and Causal Methods

Prof. Fradkin
Topic: Heterogeneous Treatment Effects

This Time

1. Heterogeneous Treatment Effects
   - Why do we want to know these?
   - The conditional average treatment effect (CATE)
   - Estimating CATE with Regression
2. The Danger of P-hacking
   - What is P-hacking
   - What is the reproducibility crisis?
   - Partial Solutions

Effects differ across people

- People vary in:

  - Interest in the product

  - Receptivity to advertising

  - Price sensitivity

  - Propensity to check their notifications

- This means that the best treatment for one person is often not the best treatment for another person.

The Conditional Average Treatment Effect (CATE)

- The conditional average treatment effect (CATE), is the effect of the treatment conditional on some characteristics.
- In the potential outcomes table on the right, the CATE for men is 0 and for women the CATE is -2/3.
- Would imply, women benefit from Zoom but not men.

Person	In-person	Zoom	Treatment Effect
John	1	1	0
Mary	0	1	-1
Suraj	0	0	0
Katerina	1	1	0
Molly	0	1	-1
Leroy	0	0	0
Average	1/3	2/3	-1/3

Estimating the CATE: Using averages

- If we are conditioning on a binary variable (such as men vs women).
- Option 1: Split the sample by the binary variable. Do a separate regression or t.test for each sample.

$$\widehat{CATE}_{men} = \bar{Y}_{men}(1) - \bar{Y}_{men}(0)$$

- In words, this is the difference between the average outcomes for treated men and control men.

Estimating the CATE: Using regression

$$Outcome_i = a + bT_i + cX_i + dX_iT_i + \epsilon_i$$

Use '*' to indicate interaction term.

```
reg_robust = sm.OLS.from_formula('racism_scores_post_2mon ~ any_treatment', data = tweets_data).fit
(cov_type='HC1')
reg_robust_hetero = sm.OLS.from_formula('racism_scores_post_2mon ~ any_treatment*anonymity', data =
tweets_data).fit(cov_type='HC1')

# print(reg_robust.summary())

result = Stargazer([reg_robust, reg_robust_hetero])
result
```

✓ 0.0s Python

Estimating the CATE: Using regression

$$Outcome_i = a + bT_i + cX_i + dX_iT_i + \epsilon_i$$

Use '*' to indicate interaction term.

```
tweets_data['anonymity_binary'] = (tweets_data['anonymity']!=0).
astype(int)
reg_robust = sm.OLS.from_formula('racism_scores_post_2mon ~
any_treatment', data = tweets_data).fit(cov_type='HC1')
reg_robust_hetero = sm.OLS.from_formula
('racism_scores_post_2mon ~ any_treatment*anonymity_binary',
data = tweets_data).fit(cov_type='HC1')

result = Stargazer([reg_robust, reg_robust_hetero])
result
```

Interpreting the CATE regression

- This regression embeds the CATE for twitter accounts that are not anonymous (anonymity = 0) and those that are anonymous (anonymity = 1).
- To get the CATE for non-anonymous: plug in 0 for ‘anonymity’ and 1 for ‘any_treatment’
- $\widehat{CATE}$ = -(.006 + .104) - .104 = -.006

	Dependent variable: racism_scores_post_2mon	
	(1)	(2)
Intercept	0.252***	0.104***
	(0.063)	(0.040)
anonymity_binary		0.172**
		(0.083)
any_treatment	−0.083	−0.006
	(0.069)	(0.052)
any_treatment:anonymity_binary		−0.092
		(0.094)
Observations	243	243
R^2	0.008	0.016

Interpreting the CATE regression

- To get the CATE estimate for anonymous: plug in 1 for ‘anonymity’ and 1 for ‘any_treatment’
- $\widehat{CATE}$ = (-.006 - .092) = -0.098

	Dependent variable: racism_scores_post_2mon	
	(1)	(2)
Intercept	0.252***	0.104***
	(0.063)	(0.040)
anonymity_binary		0.172**
		(0.083)
any_treatment	−0.083	−0.006
	(0.069)	(0.052)
any_treatment:anonymity_binary		−0.092
		(0.094)
Observations	243	243
R^2	0.008	0.016

Estimating the CATE: Continuous Covariates

$$Outcome_i = a + bT_i + cX_i + dX_iT_i + \epsilon_i$$

- Let X be a continuous variable, such as prior racism.

```
reg_robust = sm.OLS.from_formula('racism_scores_post_2mon ~
any_treatment', data = tweets_data).fit(cov_type='HC1')
reg_robust_hetero_preracism = sm.OLS.from_formula
('racism_scores_post_2mon ~
any_treatment*racism_scores_pre_2mon', data = tweets_data).fit
(cov_type='HC1')

result = Stargazer([reg_robust, reg_robust_hetero_preracism])
result
```

Estimating the CATE: Continuous Covariates

- Let's say we wanted to predict the CATE for those with a racism score of .1.
- (-.014 + .1*.309) = 0.0169
- Assumption we're making: treatment effect is linear in pre-treatment racism.

	Dependent variable: Racism (2 months after)
	(1)
Treatment	-0.014
	(0.041)
Prior Racism	0.715***
	(0.135)
Treatment*Prior Racism	0.309
	(0.440)
Constant	0.089***
	(0.032)
Observations	242
R^2	0.282
Note:	*p<0.1; **p<0.05; ***p<0.01

This Time

1. Heterogeneous Treatment Effects
   - Why do we want to know these?
   - The conditional average treatment effect (CATE)
   - Estimating CATE with Regression

2. **The Danger of P-hacking**
   - **What is P-hacking**
   - **What is the reproducibility crisis?**
   - **Partial Solutions**

Multiple-Hypothesis Testing, P-hacking, False Discovery.

FOOD FOR THOUGHT

Cornell Food Researcher's Downfall Raises Larger Questions For Science

September 26, 2018 · 3:07 PM ET

BRETT DAHLBERG

More social science studies just failed to replicate. Here's why this is good.

What scientists learn from failed replications: how to do better science.

By Brian Resnick | @B_resnick | brian@vox.com | Aug 27, 2018, 11:00am EDT

MARKETING

Is There a Replication Crisis in Research?

P-hacking

- p-hacking: the process of modifying statistical models until something is significant.
- Problem: just by chance something will be significant.
- Definition of p-val: just by chance 5% of the time, something is significant at a 5% level (pval < .05).
- https://projects.fivethirtyeight.com/p-hacking/

P-hacking

- Recipe (don’t do this!!!). Consider every variable you have (e.g. demographics, location, previous behavior) and interact various combinations with the treatment in a regression.
- After finding an effect, it’s easy to think of a theory that supports heterogeneous effects!
- Especially worrisome when the number of observations is small. Results in studies that read like this:

In two studies with large and diverse samples, ovulation had drastically different effects on single women and women in committed relationships. Ovulation led single women to become more liberal, less religious, and more likely to vote for Barack Obama. In contrast, ovulation led women in committed relationships to become more conservative, more religious, and more likely to vote for Mitt Romney.

The P-hacking smell

In two studies with large and diverse samples, ovulation had drastically different effects on single women and women in committed relationships. Ovulation led single women to become more liberal, less religious, and more likely to vote for Barack Obama. In contrast, ovulation led women in committed relationships to become more conservative, more religious, and more likely to vote for Mitt Romney.

- No main effect and no theory associated with results.
- Num. Observations < 100 for each group.
- Weird sample selection.
- (Exclude women in cycle days 1 - 6 and 26 - 28)
- Many outcome variables.
- From prior studies: Political attitudes are VERY hard to move.

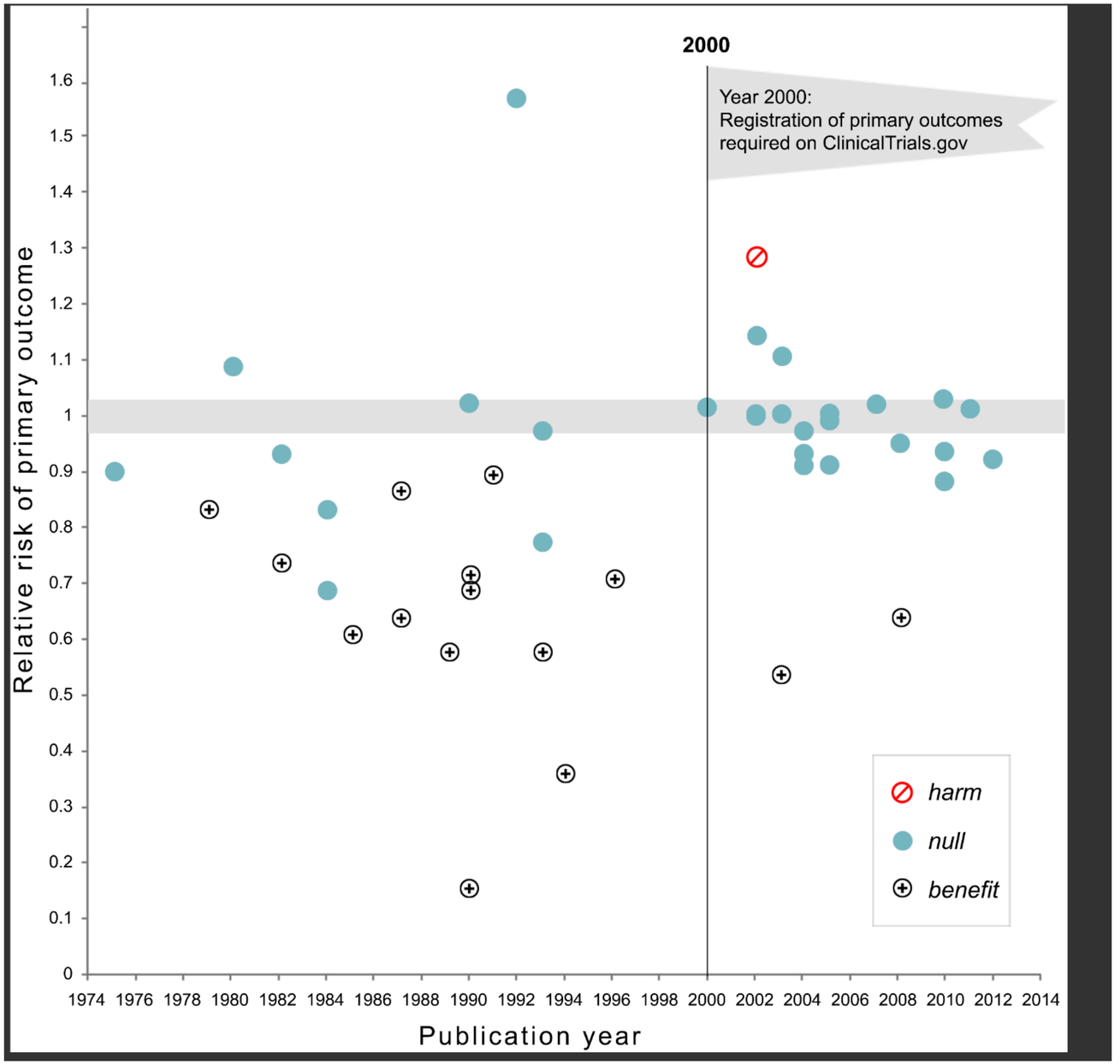
Publication Filter

- Studies with statistically significant results are much more likely to be reported and published.
- 10 people could've done the same study, but the person who by chance happened to get statistically significant effect got published.
- Note: precise 0's are also very informative! Always keep track of experiments where there is no effect.

Pre-registration

- Before running an experiment, provide a detailed plan of analysis. What methods will you use, which variables will you look at, etc…
- This reduces the worry of p-hacking.
- Power calculations should be done before the experiment is run!

- In 2000, National Heart Lung, and Blood Institute required registration of RCT on clinicaltrials.gov. More academic fields are requiring pre-registration.
- Plot shows that number of nulls increased afterward.



Recap

1. Heterogeneous Treatment Effects
   - Why do we want to know these?
   - The conditional average treatment effect (CATE)
   - Estimating CATE with Regression
2. The Danger of P-hacking
   - What is P-hacking
   - What is the reproducibility crisis?
   - Partial Solutions